Handling Imbalanced Data in Medical Records Using Entropy with Minkowski Distance

Lastri Widya Astuti[1], Ermatita[2*], Dian Palupi Rini[3]
Doctoral Program in Engineering Science, Universitas Sriwijaya, Palembang, Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia[2, 3]
Faculty of Computer & Science, Universitas Indo Global Mandiri, Palembang, Indonesia[1]

***Abstract*—Medical records are essential for disease detection to help establish a diagnosis. Many issues with imbalanced classification are discovered in many cases of early disease detection and diagnosis using machine learning methods, resulting in decreased accuracy values due to imbalanced data distribution caused by the number of positive patients with less disease than normal individuals. To improve the accuracy of the results, a classification architectural model is proposed through a modified oversampling method (SMOTE) using Minkowski distance and adding entropy as a weight value estimation to figure out the number of samples to be made. The feature selection procedure will adopt the hybrid Particle Swarm Optimisation Grey-Wolf Optimisation approach (PSO GWO). Dataset selection evaluated high, medium, and low data dimensions based on the number of features and the total number of dataset samples. The six classification algorithms were compared using datasets involving diabetes, heart, and breast cancer. The final classification results indicated an average accuracy of 74% for diabetes, 83% for heart, and 96% for breast cancer. The proposed approach successfully solves imbalances in medical record data, outperforming Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest classification approaches.**

***Keywords*—Medical record; imbalanced data; classification; distance; entropy**

I. INTRODUCTION

Developing professional and quality health services requires a knowledge base to provide health services for patients based on evidence of the latest medical actions obtained from medical records of every action performed on patients [1]. Medical records are considered a key factor for improving the quality and safety of health services [2]. The use of medical records as documentation of patient care is increasingly being used to understand the clinical manifestations of various diseases, as well as the relationship between multiple diseases [3] [4]. Different knowledge bases have been engineered to extract useful information from medical records to assist the learning process for health professionals in making diagnoses [5] [6].

Medical record data stores many critical patient attributes from several medical action records and patient health records that describe a health condition [7] [8]. In many cases of early disease detection and diagnosis using machine learning methods, the problem of unbalanced classification is a case that is often encountered where the distribution of negative patients is greater than that of positive patients with the disease, causing an imbalance [9]. Several previous studies have shown that imbalanced data causes a decrease in accuracy values due to unequal class distribution, causing minority classes to be grouped into the majority class [10].

Several improvement methods are proposed to balance the amount of data in different classes through integration methods or improving learning methods by adding samples to the minority class (oversampling) or by removing samples from the majority class (undersampling) [11]. Unbalanced algorithms have three learning approach methods: data level, algorithm level, and hybrid methods [12]. The data level method improves classification performance by changing the distribution of data sets based on classes to balance the data. In contrast, the algorithm level method modifies the classification algorithm to adapt to unbalanced data structures, while the hybrid method combines the data level method and the algorithm level method in an integrated manner. Simultaneously, it is necessary to achieve the best performance [5] [13].

Oversampling techniques have been proven effective in resolving data imbalance problems. The Synthetic Minority Oversampling Technique (SMOTE) is the most popular technique for overcoming data imbalance [14]. The SMOTE method interpolates several data points from the minority class by considering neighbouring point data to produce new samples, but the interpolation process is ineffective for high-dimensional data [15] [16]. SMOTE uses Euclidean distance to calculate the neighbour distance for each minority class instance. They often increase overlap because the distance between each other is almost uniform and converges to the same value for all the cases, thereby reducing the accuracy value during the classification process [17]. Several previous studies have made improvements to the SMOTE method, including adding weights [18], learning rate [19], or combining two algorithms to improve performance and accuracy [20]. This research proposes modifications to SMOTE by using Minkowski distance and adding entropy as a weight value calculation to calculate the number of samples to be made. Minkowski distance is a measurement metric in normed vector spaces built on the distance spectrum and is considered a generalization of Euclidean and Manhattan distance. Minkowski distance can also describe the geometric structure of majority and minority class data [21]. The plan being developed for ensemble models will integrate the balanced data and then select the best features to be utilized by the classification method to achieve the highest level of accuracy. All stages will be carried out in a structured manner following the proposed design, and the research contribution will be determined so that the direction is clear and measurable.

The following are the contributions to this research:

1) This research presents improvements to the SMOTE method to overcome data imbalance by adding entropy as a weight to generate the number of data instances and changing the distance learning method to overcome overfitting.

2) The feature selection process in this research presents a hybrid architecture that combines Particle Swarm Optimization (PSO) with Grey Wolf Optimization (GWO). Combining two algorithms for feature selection can reduce the number of attributes to increase accuracy and computing time.

3) This research compares several classification methods, such as Decision Tree, K-NN, Logistic Regression, Naïve Bayes, Random Forest, and SVM, to see the best performance of the six methods used in disease classification research experiments.

4) Ensemble learning techniques improve medical record data classification results in this research. This research aims to optimize the accuracy of classification results by using the advantages of various model comparisons to overcome the complexity of data imbalance and reduce data dimensions. That offers a perspective for improving the accuracy of pattern recognition results in medical record data.

II. Literature Review

A. Imbalance Data

Class imbalance is a condition where the number of *instances of* the majority class is greater than the number *of cases of the* minority class. The difference in the number of samples in each class in the classification is known as an unbalanced data set [22]. Extension *ensemble* for unbalanced learning considers changes in the pattern probability function during the training phase, where training on unbalanced data tends to overtrain the majority class, while training for the minority class becomes difficult to predict due to undertraining [23][24]. Data imbalance can be divided into two types, namely relative imbalance and absolute imbalance, while based on the cause of the problem, the imbalance can be divided into two: intrinsic and extrinsic. Intrinsic indicates an inherent property of the data. For example, the probability of equipment failing is much lower than standard running equipment, or the number of people with cancer is less than that of healthy people. Extrinsic means other or external factors that cause data imbalance. For example, sporadic interruptions occur when balanced data is transmitted to the database [25].

B. Feature Selection

Dimensionality reduction is done by selecting features to increase accuracy by eliminating irrelevant features and assigning selected features to make the data easier to understand [26]. In the feature determination process, there are three approaches used to select features:

1) Filter method, selecting features based on feature relevance using ranking criteria, where the lowest calculated value will be removed. The filter method does not depend on the classification algorithm used in modelling so that it can be generalized and impacts accelerating computing time.

2) The wrapper method is part of a supervised algorithm where features are selected based on the relationship between features by comparing the resulting subset of features. The training and testing process is based on predicting the representativeness of each feature, which causes the computing time to become more complex.

3) The embedded method performs feature selection by looking for an optimal subset of features and including them in the training process of the classifier, thereby reducing computing time.

C. Classification

Utilization of machine learning helping to enforce fast, accurate diagnosis and detection is one of the efforts made in the health sector. Algorithm development machine learning, used to predict and diagnose disease based on medical record data, can help reduce diagnosis errors and independent pre-diagnosis learning materials. Several studies are related to solving or detecting diseases through classification methods, including the K-NN method, which is used to classify medical health data related to diseases and drugs [27] [28]. Method Naïve Bayes and Random Forest, which is used to diagnose and detect various diseases [29] [30], the SVM method is used to classify medical record data in the form of images [31] [32]. The utilization of machine learning also touches on digital data-based disease classification data processing, such as the utilization of telehealth and IoT using the decision tree method [33]. Combining two machine learning methods or modifying a technique has been developed to improve medical record data classification results and diagnose diseases [34]. The classification results of medical record data in previous research show variations in accuracy. Several factors influence the level of accuracy, including high-dimensional data and data imbalance.

The model proposed in this research is to overcome data imbalance in medical record data, where in several previous studies, many modifications or additional functions have been made to overcome the imbalance problem by adding entropy and changing the distance calculation to make it more flexible for data with different dimensions. Meanwhile, the feature selection process uses a metaheuristic method that focuses on finding optimal solutions flexibly. PSOGWO is used to optimize subsets to get features that provide the best performance on the model. This model was tested using six algorithms: decision tree, logistic regression, naïve Bayes, KNN, random forest, and SVM to test the ensemble combination that produces the most optimal level of accuracy.

III. Methodology

The methodology used in the research contains comprehensive stages of each built model process. The stages for overcoming the data imbalance problem begin with preparing a medical record dataset, data preprocessing, feature selection, classification, and evaluation methods. The research stages are shown in Fig. 1.

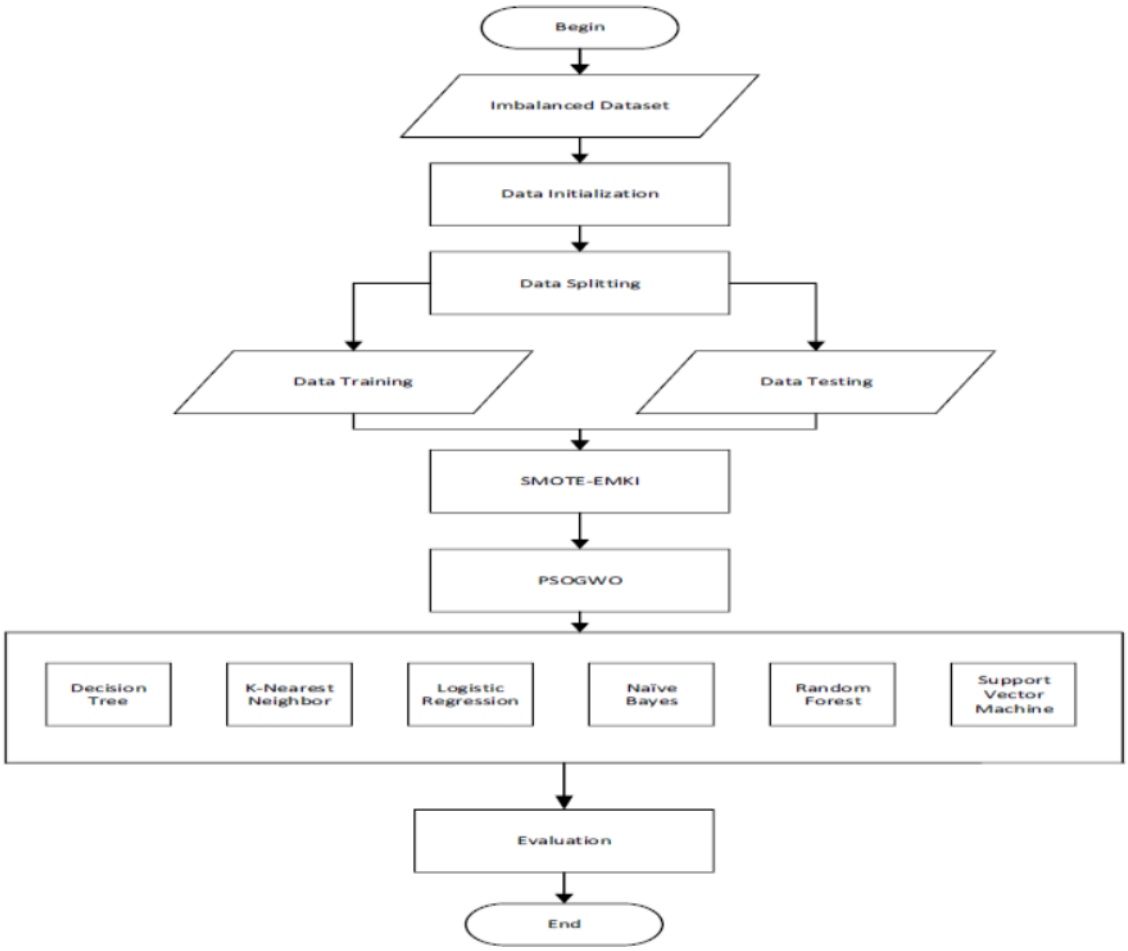

Fig. 1. The proposed architecture.

A. Dataset

The dataset used in this research is medical record data for diabetes, heart disease, and breast cancer, accessed in the UC repository *machine learning* and Kaggle. The selection of medical record data is based on the prevalence rate of sufferers, which increases yearly. Detailed data consists of the number of features, number of example data, number of minor data, number of classes, and Imbalanced Ratio (IR), described in Table I.

TABLE I. DESCRIPTION OF MEDICAL RECORDS DATASET

Dataset	Feature	Instance	Minority	Class	IR
Diabetes	8	768	268	2	1.86
Heart	14	1025	499	2	1.054
Breast Cancer	32	569	212	2	1.68

B. Data Imbalance Method

The approach used in this research is to modify the SMOTE algorithm by adding an entropy value at the beginning of the data balancing process and changing the distance calculation method from Euclidean Distance to Minkowski with the advantage of distance adjustment capabilities. The development stages of this model start from:

Step 1: Calculate the entropy weight for each majority and minority class. Entropy is used to measure how balanced the class distribution is in the dataset, where the entropy calculation for each class is carried out before the distance calculation. The entropy equation used is:

$$Entropy\ (S) = \ \sum_{i=1}^{C} p_{i\ log_2}(p_i) \qquad (1)$$

High entropy values indicate a more significant imbalance, while low entropy values indicate a more balanced class distribution.

Step 2: Determine the number of synthetic samples. The number of synthetic samples that need to be added is calculated based on the entropy proportion of the majority and minority

classes. If the entropy of the minority class is higher, the class imbalance is more significant, so many synthetic samples need to be added.

Step 3: Calculate the distance according to the given weight. After calculating the entropy value, normalization is carried out before proceeding to the distance calculation. Where the previous distance calculation using Euclidean distance was changed using Minkowski to calculate the closest K value from the minority class with the equation:

$$D_{(p,q)} = \left(\sum_{i=1}^{n} |p_i - q_i|^r\right)^{1/r} \quad (2)$$

Step 4: Synthetic interpolation creates synthetic samples based on linear interpolation between minority data points and their nearest neighbours.

One of the problems in SMOTE is that the synthetic samples produced have a high level of similarity to existing samples, so they do not adequately represent the diversity of minority classes. Adding entropy in the SMOTE technique is intended to improve the quality of synthetic samples and reduce the possibility of bias. Entropy is used to help select sample points and interpolation, where the results can increase the representativeness and variation of synthetic samples. Entropy positioning is also a consideration in this research; entropy is placed before applying the SMOTE method with consideration of the entropy value used to determine the number of synthetic samples needed. The entropy value also helps assess relevant features and eliminates the possibility of redundant minority classes. Placing entropy before the SMOTE process helps increase data diversity, accelerates convergence, and reduces the potential for overfitting.

A part from adding entropy, the proposed model also changes the distance calculation technique from Euclidean Distance to Minkowski Distance. This change aims to provide flexibility in measuring the distance between data points for data with different dimensions, thereby increasing the quality of synthetic samples. Changing the distance calculation method to Minkowski distance also helps determine the nearest neighbour selection of parameters to produce synthetic samples. In addition, the parameters can be adjusted to obtain distance measurements that better suit the characteristics of the data or specific features so that when the parameters are chosen appropriately, the diversity of the resulting synthetic samples can be increased. The SMOTE algorithm has been modified by adding entropy and replacing the distance approach with Minkowski, now called EMKI SMOTE. Furthermore, the results of improving this method will be used in the classification process, which was previously preceded by a feature selection process.

C. Algorithm PSO GWO

One of the population-based metaheuristic optimization techniques is Particle Swarm Optimization (PSO), where the initial inspiration for this method was the social behavior of flocks of birds or schools of fish when foraging for food [35] [36]. In this technique, a number of particles are allowed to move in a multidimensional search space, where an initial population is generated randomly in the search domain [37] [38]. All particles in the moving swarm update their positions using Eq. (3) and Eq. (4) at each iteration. The entire change history of the best position information of each particle in the cluster is also updated and saved.

$$X^{-i}{}_{n+1} = X_n^{-1} + V_{n+1}^{-1}, \quad (3)$$

$$V^{-i}{}_{n+1} = \omega V^{-1}{}_n + c_1 r_1 \left(\overline{p}^i{}_n - \overline{x}^i{}_n\right) \quad (4)$$

The particle symbols used in swarm for optimization are as follows:

N: The iteration steps carried out

r1 dan r2: Value representing a random number in the range [0, 1]

ω : Parameter weight inertia

c1 dan c2 : Coefficients representing optimization parameters

x : Position vector

v : Velocity vector

$\overline{x}^i$: The best position that the i particle has achieved

$\overline{p}^g$: The best position representation available on the *swarm*.

In the PSO algorithm, the position and velocity of the particles are determined randomly in the search space to find the best value. The goal of this operation is to avoid local minima. The search is continued until optimal results are achieved or based on achieving a predetermined maximum number of iterations. The iteration process is carried out to obtain the best solution through the following equation:

$$x_i^{k+1} = \begin{Bmatrix} x_i new^{i \iint (x_i new) \leq \int (x_i)} \\ x_i otherwise \end{Bmatrix} \quad (5)$$

The grey wolf leadership hierarchy at the top of the food chain is the rationale for the GWO algorithm. The grey wolf group is divided into four categories, alpha, beta, delta and omega. The alpha wolf represents the highest hierarchy, which provides the best solution for the group, while the second and third hierarchies are occupied by the beta and delta wolves, respectively. The omega wolf's role in the pack as the best solution candidate in the pack if needed. The stages in hunting prey by groups of grey wolves are divided into three parts, namely: the first stage is carrying out reconnaissance on the prey to be hunted, the second stage is stopping the movement of the prey by circling and confining it so that it makes it easier to attack and prey on the victim as the final stage. Eq. (6) and Eq. (7) provide a mathematical model for surrounding the victim.

$$D = \left|C \times X_p(t) - X(t)\right| \quad (6)$$

$$X(t+1) = X_p(t) - A \times D \quad (7)$$

Here, t denotes the number of instantaneous iterations, X_p is the prey position, X is the grey wolf's location, and A and C

are the vector coefficients. The coefficients A and C are calculated as shown in Eq. (8) and Eq. (9).

$$A = a \times (2 \times r_1 - 1) \quad (8)$$

$$C = 2 \times r_2 \quad (9)$$

The number of drops linearly from two to zero as the number of iterations decreases. r_1 and r_2 are random numbers picked uniformly from [0,1].

Alpha wolves direct grey wolves to the location of their prey. Alpha wolves occasionally receive assistance from beta and delta wolves. The GWO method assumes that the alpha wolf is the best option, with beta and delta wolves coming in second and third. As a result, the other wolves in the population migrated following the positions of these three wolves. The formulated mathematically in Eq. (10), (11), and (12):

$$D_\alpha = |C_1 \times X_\alpha - X(t)| \quad (10)$$

$$D_\beta = |C_2 \times X_\beta - X(t)| \quad (11)$$

$$D_\delta = |C_3 \times X_\delta - X(t)| \quad (12)$$

Values X_α, X_β, and X_δ each represent the three best wolves in each iteration.

$$X_1 = |X_\alpha - a_1 D_\alpha| \quad (13)$$

$$X_2 = |X_\beta - a_2 D_\beta| \quad (14)$$

$$X_3 = |X_\delta - a_3 D_\delta| \quad (15)$$

$$X_P(t+1) = \frac{X_1 + X_2 + X_3}{3} \quad (16)$$

Here, the new prey position is expressed as Xp(t + 1) as the mean of the positions of the three best wolves in the population. The grey wolf completes the hunt by attacking its prey. To attack, they must be close enough to their prey. When Eq. (5) is checked, A takes varying values from [− 2a, 2a] while A takes decreasing values from 2 to 0. When the |A| value exceeds or equals 1, the existing hunt is abandoned to find a better solution. The grey wolf is forced to attack the prey if the prey is close enough to a value less than 1. This approach prevents wolves from getting stuck at local minima. The search is complete when the GWO algorithm reaches the desired number of iterations.

D. Evaluation Metric

An evaluation of model performance is needed to see the effectiveness of the proposed ensemble model. Model performance is measured using accuracy, precision, recall, and F1 score metrics based on actual and predicted results from the classification process, where the basis for calculating metrics can be defined as follows: True Positive (TP) notation: the amount of data labelled positive and classified by the model labelled positive; True Negative (TN) is the amount of data labelled negative and classified as negative; False Positive (FP) is the amount of actual data labelled negative and predicted positive; False Negative (FN) is the amount of actual data labelled positive and predicted negative.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

$$F1\ score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (20)$$

IV. Result and Discussion

The designed EMKI SMOTE model causes the process of creating synthetic samples in the minority class to become more dynamic and diverse. The modification results in Table II show a significant improvement in overcoming data imbalance. The imbalance ratio (IR) in the Table compares the ratios before and after the modification was applied.

TABLE II. Imbalance Ratio

Dataset	Imbalance Ratio	
	SMOTE	EMKI SMOTE
Diabetes	1: 1.86	1: 1.47
Heart	1: 1.05	1: 1.02
Breast Cancer	1: 1.68	1: 1.43

Applying the EMKI SMOTE method to the model allows for the balance of data and the production of diverse synthetic samples so that the minority class data is more representative. Adding entropy and the Minkowski method for calculating distance can also eliminate redundant sample selection, reducing the risk of overfitting. After completing the data imbalance stage, the model will continue with the feature selection stage.

Particle Swarm Optimization - Grey Wolf Optimization (PSO-GWO) feature selection is a hybrid method that takes advantage of the benefits of both optimization algorithms to pick the most relevant subset of features in a dataset. PSO (Particle Swarm Optimization) is an algorithm inspired by the behaviour of flocks of birds or fish, in which each particle represents a solution and interacts with the others to discover the optimal position. Meanwhile, GWO (Grey Wolf Optimization) is an algorithm inspired by grey wolf hunting behaviours that uses a highly efficient search mechanism that divides roles among alpha, beta, and delta wolves. In feature selection, PSO identifies an initial solution, and GWO enhances the search for the solution by steering it in a more efficient path. The combination of these two algorithms aids in overcoming the dataset's complexity and feature redundancy, resulting in a more efficient model with improved performance. The PSO-GWO method can avoid traps in local search and speed up feature selection while maintaining model fidelity.

Feature selection in this study uses a combined algorithm between Particle Swarm Optimization and Grey Wolf Optimization (PSO-GWO) to produce the best feature selection process from the tested medical record datasets. The selection of features using this method will select attributes as representatives for each dataset. Fig. 2 shows the results of feature selection before using the feature selection method, after using the PSO GWO method, and the results of feature selection after the dataset is balanced, showing the results of feature selection with the optimal number of features without losing information for the classification process.

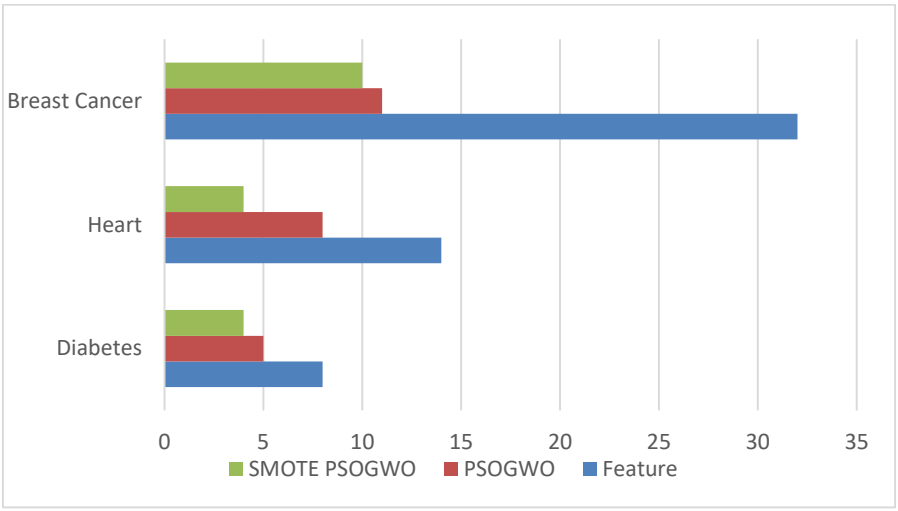

Fig. 2. Feature selection results using the PSO GWO.

The classification process divides the dataset into two parts: training data and testing data. The data is divided proportionally using the holdout method. Training data helps the model recognize patterns or relationships between features and labels while testing data measures the accuracy or significance of the model in classifying data that has not been introduced previously.

The experiment used medical record data for diabetes, heart disease, and breast cancer accessed from the UCI Machine Learning Repository. The results of feature selection with balanced data then go through the classification stage. The classification process uses a Decision Tree, KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM methods. The classification method comparison process is intended to strengthen evidence regarding the performance of the proposed model. Fig. 3 shows a confusion matrix graph for testing data on diabetes.

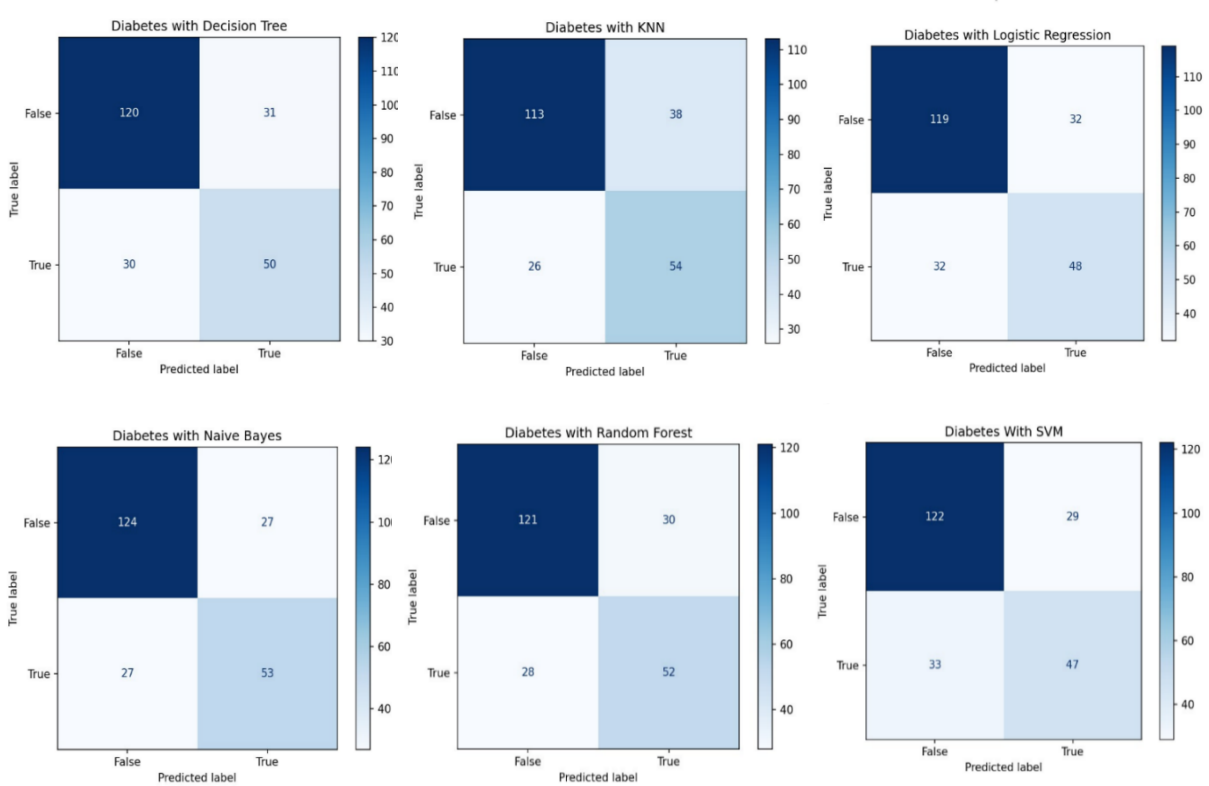

Fig. 3. Confusion matrix for the diabetes dataset.

The test results show that the diabetes dataset, when tested using six classification methods, shows an average Accuracy value of 0.7411, with an average value for precision of 0.79355, while an average value for recall is 0.8026 and an average F1 score of 0.7984. The methods that provide the best performance results are the Naïve Bayes and Random Forest methods, while the lowest performance results in experiments using the Decision Tree method. Details of the experimental results using each classification method are shown in Table III.

TABLE III. EVALUATION MATRIX DIABETES

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Diabetes	Decision Tree	0.7359	0.7947	0.7947	0.7946
	K Nearest Neighbor	0.7229	0.7483	0.8129	0.7792
	Logistic Regression	0.7229	0.7880	0.7880	0.7887
	Naïve Bayes	0.7662	0.8211	0.8211	0.8210
	Random Forest	0.7489	0.8013	0.8120	0.8066
	Support Vector Machine	0.7316	0.8079	0.7870	0.8008

The second test was carried out on the heart disease dataset, where the classification process was given the same treatment for each stage as in the diabetes dataset. The confusion matrix of the results of heart disease data testing is shown in Fig. 4. The experimental results show increased accuracy, precision, recall, and F1 score after the medical data was balanced using EMKI SMOTE. Analysis of the experimental results also shows that the PSO GWO method used in the feature selection process combined with EMKI SMOTE produces several relevant features to use during the classification process.

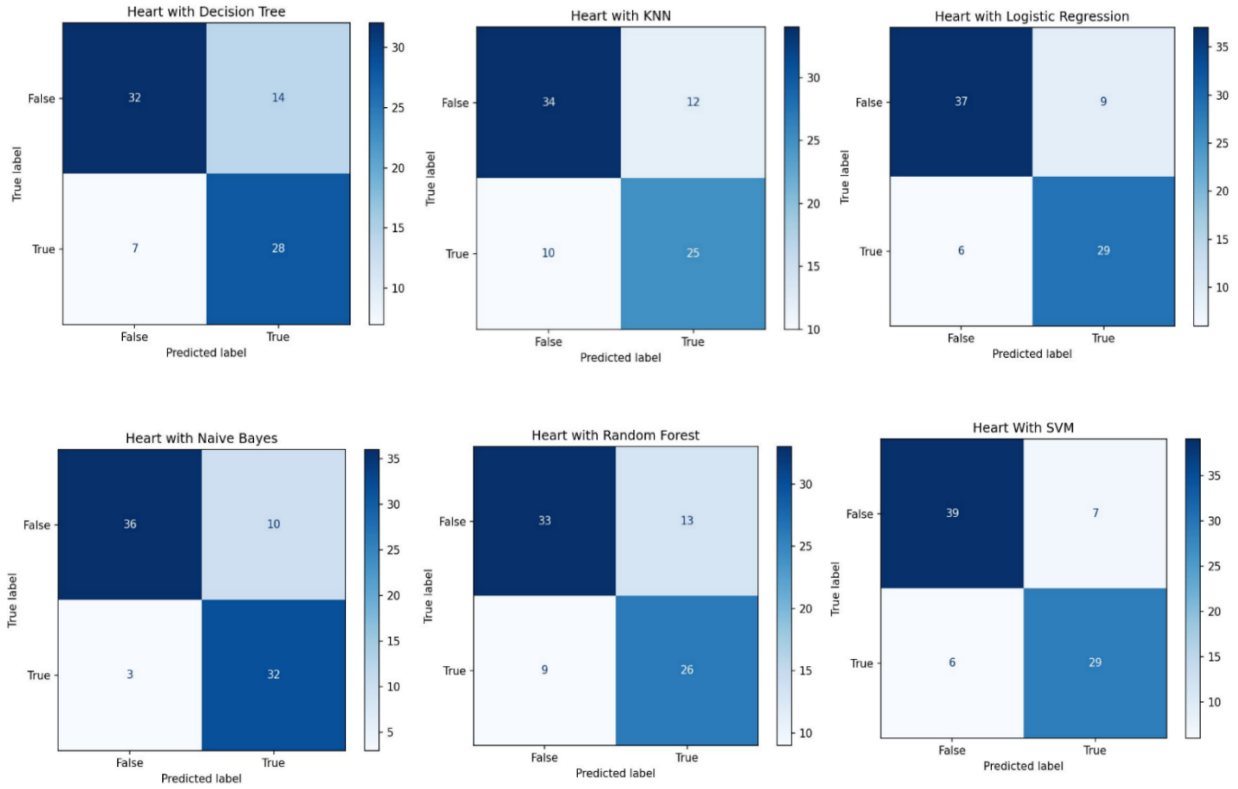

Fig. 4. Confusion matrix for heart disease data.

Experiments using the heart disease dataset showed an average accuracy of 0.7818, precision of 0.7644, recall of 0.8381, and F1 Score of 0.7989. Meanwhile, the best-combined method is EMKI SMOTE, PSO GWO combined with SVM, Naïve Bayes, and Logistic Regression, which is shown as a whole in Table IV.

The third experiment was carried out on the breast cancer dataset, with 32 features. The experimental results of the model also show that the feature selection results are significant and relevant, thus influencing the improvement of experimental results. The Minority class value is lower than the majority, with an imbalance ratio of 1.68, after being corrected with the proposed model, indicating good and even experimental results for the six classification methods.

TABLE IV. EVALUATION MATRIX HEART

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Heart	Decision Tree	0.7407	0.6956	0.8205	0.7527
	K Nearest Neighbor	0.7283	0.7391	0.7727	0.7555
	Logistic Regression	0.8148	0.8043	0.8604	0.8313
	Naïve Bayes	0.8395	0.7826	0.9230	0.8469
	Random Forest	0.7283	0.7173	0.7857	0.7499
	Support Vector Machine	0.8395	0.8478	0.8667	0.8571

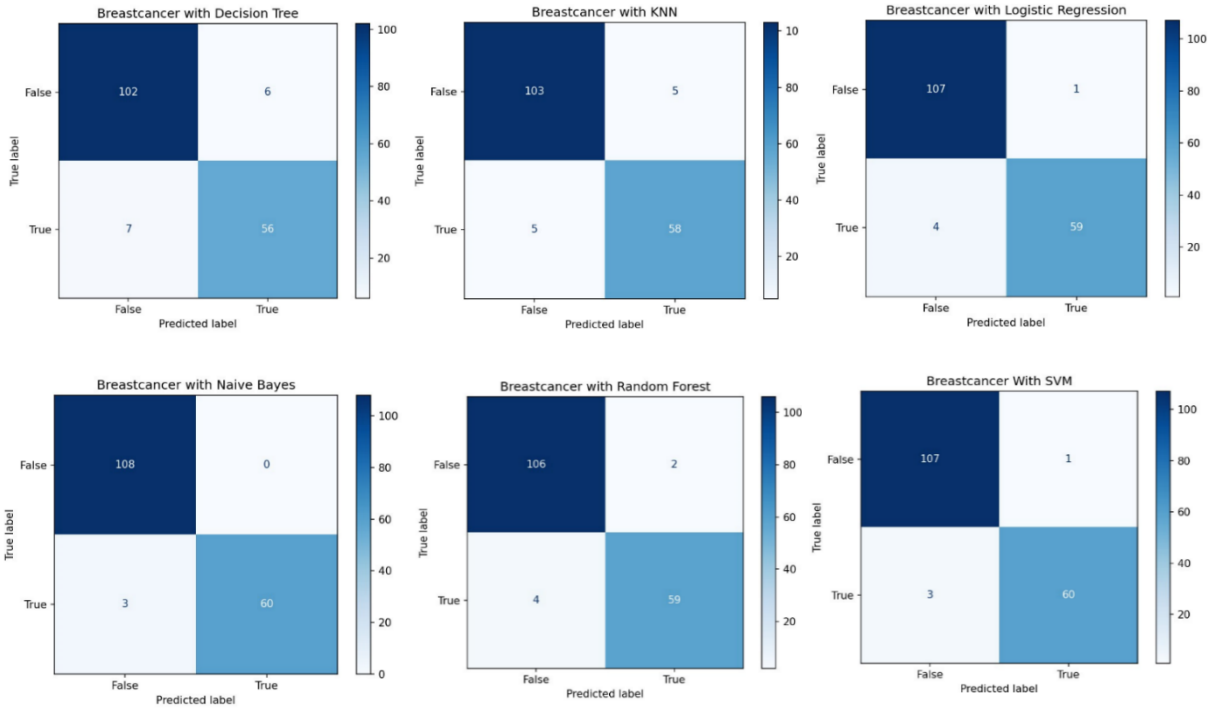

Fig. 5. Confusion matrix for breast cancer.

The application of the EMKI SMOTE model combined with PSO GWO produces averages for accuracy, precision, recall, and F1 Score in the range of 0.9684 with the best results in the Naïve Bayes, SVM, and Logistic Regression classification methods as shown in Table V.

The application of the EMKI SMOTE model combined with the PSO GWO method on medical record data was due to the discovery of a lot of unbalanced data, with the prevalence of patients in normal conditions being higher than patients suffering from disease. The experiment also selected datasets by considering high, medium, and low data dimensions based on the number of features and the total number of dataset samples. Analysis of experimental results from various aspects shows a significant increase in results, and computing time calculations also indicate that the training and testing process is faster.

TABLE V. EVALUATION MATRIX BREAST CANCER

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Breast Cancer	Decision Tree	0.9239	0.9444	0.9357	0.9399
	K Nearest Neighbor	0.9415	0.9537	0.9537	0.9536
	Logistic Regression	0.9707	0.9907	0.9639	0.9770
	Naïve Bayes	0.9824	1.00	0.9729	0.9863
	Random Forest	0.9649	0.9814	0.9636	0.9723
	Support Vector Machine	0.9766	0.9907	0.9727	0.9816

V. CONCLUSION

Imbalanced class data has become an essential problem in medical diagnosis, but using classification algorithms alone cannot meet the classification needs effectively. This research proposes a combined method by modifying the distance calculation using Minkowski and adding entropy to determine the value of the number of samples created. The proposed calculation model is hereafter called EMKI SMOTE. The EMKI SMOTE oversampling approach is designed to handle imbalances between majority and minority classes in data so that it can produce relevant features in high-dimensional data. Next, the balanced data will be ensembled using the PSO GWO method in the feature selection process. The combined results of the data balance method and feature selection will then be used in the training and testing process using the classification method. The research results show increased accuracy, especially in the Naïve Bayes method, Logistic Regression, and SVM using diabetes, heart, and breast cancer datasets. The further development of this research model is to obtain constant balanced data, namely, by determining the interval between the majority and minority sample numbers. In addition to adding parameters to the dataset model configuration, it is possible to balance the data, which causes an increase in accuracy. Meanwhile, testing large amounts of data and high dimensions by making several substitutions in the feature selection method is also considered to increase the accuracy of classification results and help in the problem of early disease detection.

ACKNOWLEDGMENT

We are grateful to everyone who helped with this study, especially the promoter and co-promoter who advised completing it. In addition, I want to convey my heartfelt gratitude to the Indo Global Mandiri University for allowing me to improve personally.

REFERENCES

[1] Y. Huang, W. Bian, and Y. Han, "Effect of knowledge acquisition on gravida's anxiety during COVID-19," Sex. Reprod. Healthc., vol. 30, no. 639, p. 100667, 2021, doi: 10.1016/j.srhc.2021.100667.

[2] A. De Benedictis, E. Lettieri, L. Gastaldi, C. Masella, A. Urgu, and D. Tartaglini, "Electronic medical records implementation in hospital: An empirical investigation of individual and organizational determinants," PLoS One, vol. 15, no. 6, pp. 1–12, 2020, doi: 10.1371/journal.pone.0234108.

[3] A. Caccamisi, L. Jørgensen, H. Dalianis, and M. Rosenlund, "Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records," Ups. J. Med. Sci., vol. 125, no. 4, pp. 316–324, 2020, doi: 10.1080/03009734.2020.1792010.

[4] H. L. Lin, D. C. Wu, S. M. Cheng, C. J. Chen, M. C. Wang, and C. A. Cheng, "Association between Electronic Medical Records and Healthcare Quality," Med. (United States), vol. 99, no. 31, 2020, doi: 10.1097/MD.0000000000021182.

[5] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," IEEE Access, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[6] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," Int. J. Med. Inform., vol. 144, no. September, p. 104302, 2020, doi: 10.1016/j.ijmedinf.2020.104302.

[7] X. Li, H. Wang, H. He, J. Du, J. Chen, and J. Wu, "Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks," BMC Bioinformatics, vol. 20, no. 1, pp. 1–12, 2019, doi: 10.1186/s12859-019-2617-8.

[8] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui, "A Survey of Deep Learning for Electronic Health Records," Appl. Sci., vol. 12, no. 22, pp. 1–24, 2022, doi: 10.3390/app122211709.

[9] W. Zhang, X. Li, X. D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," Meas. J. Int. Meas. Confed., vol. 152, 2020, doi: 10.1016/j.measurement.2019.107377.

[10] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," Int. J. Artif. Intell. Res., vol. 4, no. 2, p. 86, 2021, doi: 10.29099/ijair.v4i2.152.

[11] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "Knowledge-Based Systems LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," vol. 196, 2020, doi: 10.1016/j.knosys.2020.105845.

[12] Q. Dai, J. Liu, and Y. Liu, "Multi-granularity Relabeled Under-sampling Algorithm for Imbalanced Data," 2022.

[13] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[14] H. A. Gameng, B. B. Gerardo, and R. P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets," ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., pp. 19–23, 2019, doi: 10.1109/ICETAS48360.2019.9117287.

[15] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," Appl. Soft Comput. J., 2018, doi: 10.1016/j.asoc.2018.12.024.

[16] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE," Int. J. Comput. Intell. Syst., vol. 12, no. 2, pp. 1412–1422, 2019, doi: 10.2991/ijcis.d.191114.002.

[17] E. A. Pambudi and A. G. Fahrezi, "Forecasting Brown Sugar Production Using k-NN Minkowski Distance and Z-Score Normalization," vol. 5, no. 2, pp. 580–589, 2023, doi: 10.51519/journalisi.v5i2.485.

[18] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE : A feature-weighted oversampling approach for imbalanced classification," vol. 124, 2022, doi: 10.1016/j.patcog.2021.108511.

[19] Y. Wang, Y. Wei, H. Yang, J. Li, Y. Zhou, and Q. Wu, "Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–13, 2020, doi: 10.1186/s12911-020-01245-4.

[20] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," Sci. J. Informatics, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[21] C. Fu, "Granular Classification for Imbalanced Datasets : A Minkowski Distance-Based Method," 2021.

[22] K. Roy et al., "An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values," Complexity, vol. 2021, 2021, doi: 10.1155/2021/9953314.

[23] D. Fahrudy and S. 'uyun, "Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection," Int. J. Informatics Vis., vol. 6, no. 4, pp. 798–808, 2022, doi: 10.30630/joiv.6.4.982.

[24] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," IOP Conf. Ser. Mater. Sci. Eng., vol. 1099, no. 1, p. 012077, 2021, doi: 10.1088/1757-899x/1099/1/012077.

[25] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," J. Big Data, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0151-6.

[26] A. G. Hussien, D. Oliva, E. H. Houssein, A. A. Juan, and X. Yu, "Binary whale optimization algorithm for dimensionality reduction," Mathematics, vol. 8, no. 10, pp. 1–24, 2020, doi: 10.3390/math8101821.

[27] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," IEEE Access, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[28] F. Aldi, I. Nozomi, and S. Soeheri, "Comparison of Drug Type Classification Performance Using KNN Algorithm," SinkrOn, vol. 7, no. 3, pp. 1028–1034, 2022, doi: 10.33395/sinkron.v7i3.11487.

[29] T. Widiyaningtyas, I. A. E. Zaeni, and N. Jamilah, "Diagnosis of fever symptoms using naive bayes algorithm," ACM Int. Conf. Proceeding Ser., pp. 23–28, 2020, doi: 10.1145/3427423.3427426.

[30] J. E. Aurelia, Z. Rustam, I. Wirasati, S. Hartini, and G. S. Saragih, "Hepatitis classification using support vector machines and random forest," IAES Int. J. Artif. Intell., vol. 10, no. 2, pp. 446–451, 2021, doi: 10.11591/IJAI.V10.I2.PP446-451.

[31] J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review," IEEE Access, vol. 8, pp. 150489–150513, 2020, doi: 10.1109/ACCESS.2020.3016782.

[32] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "Lung cancer medical images classification using hybrid cnn-svm," Int. J. Adv. Intell. Informatics, vol. 7, no. 2, pp. 151–162, 2021, doi: 10.26555/ijain.v7i2.317.

[33] C. C. Chern, Y. J. Chen, and B. Hsiao, "Decision tree-based classifier in providing telehealth service," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–15, 2019, doi: 10.1186/s12911-019-0825-9.

[34] C. Iwendi et al., "COVID-19 patient health prediction using boosted random forest algorithm," Front. Public Heal., vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[35] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection," IEEE Access, vol. 7, no. c, pp. 39496–39508, 2019, doi: 10.1109/ACCESS.2019.2906757.

[36] M. A. M. Shaheen, H. M. Hasanien, and A. Alkuhayli, "A novel hybrid GWO-PSO optimization technique for optimal reactive power dispatch problem solution," Ain Shams Eng. J., vol. 12, no. 1, pp. 621–630, 2021, doi: 10.1016/j.asej.2020.07.011.

[37] S. Prithi and S. Sumathi, "Automata Based Hybrid PSO–GWO Algorithm for Secured Energy Efficient Optimal Routing in Wireless Sensor Network," Wirel. Pers. Commun., vol. 117, no. 2, pp. 545–559, 2021, doi: 10.1007/s11277-020-07882-2.

[38] Z. H. A. Al-Tameemi, T. T. Lie, G. Foo, and F. Blaabjerg, "Optimal Coordinated Control of DC Microgrid Based on Hybrid PSO–GWO Algorithm," Electricity, vol. 3, no. 3, pp. 346–364, 2022, doi: 10.3390/electricity3030019